Situation Based Speech Recognition for Structuring Baseball Live Games

Atsushi Sako, Tetsuya Takiguchi and Yasuo Ariki

Department of Computer and Systems Engineering
Kobe University, Kobe, Japan
sakoats@me.cs.scitec.kobe-u.ac.jp
{ariki,takigu}@kobe-u.ac.jp

Abstract

It is a difficult problem to recognize baseball live speech because the speech is rather fast, noisy, emotional and disfluent due to rephrasing, repetition, mistake and grammatical deviation caused by spontaneous speaking style. To solve these problems, we have been studied the speech recognition method incorporating the baseball game task-dependent knowledge as well as an announcer's emotion in commentary speech [1]. In addition, in this paper, we propose the situation prediction model based on word co-occurrence. Owing to these proposed models, speech recognition errors are effectively prevented. This method is formalized in the framework of probability theory and implemented in the conventional speech decoding (Viterbi) algorithm. The experimental results showed that the proposed approach improved the structuring and segmentation accuracy as well as keywords accuracy.

1. Introduction

Recently a large quantity of multimedia contents are broadcast and accessed through digital TV and WWW. In order to retrieve exactly what we want to know from multimedia database, automatic extraction of meta information or structuring is required, because it is impossible to give them to multimedia database manually due to their quantity.

The purpose of this study is to improve the speech recognition accuracy for automatically transcribing sports live speech, especially baseball commentary speech, in order to produce the closed caption and to structure the sports games for highlight scene retrieval. The baseball game structuring is the process of segmenting the game into a pitching sequence and giving each of them the meta information such as inning, out count, strike count and ball count. The accuracy of the baseball game structuring depends on the transcription accuracy so that sophisticated speech recognition techniques are required.

As the sports live speech, we used radio speech instead of TV speech because the radio speech has much more information about the keywords. However the radio speech is rather fast, noisy and emotional. Furthermore, it is disfluent due to rephrasing, repetition, mistake and grammatical deviation caused by spontaneous speaking style. To solve these problems, we have already proposed the adaptation techniques of language model and acoustic model, which convert a baseline model originally constructed using available speech corpus to the sports live model using sports live speech[2].

In order to further improve the speech recognition accuracy, we propose in this paper, the *Situation Based Speech Recognition* method. Here some related works are reported about situation based language modeling especially around dialogue systems[5][6][7]. In previous works, *a situation* is considered *a dialogue state*. The prediction of a dialogue state is easily since a dialogue state is defined by system's question that a user is replaying to. In our method, there are some differences from the dialogue systems. First, we consider *a structure* of baseball games — counting of inning, out, ball and strike — and *announcer's emotion* as a situation. Second, the prediction of a baseball game situation is difficult because we have to predict a situation by speech recognition results. Last, we consider not only situation dependent language models but also situation dependent acoustic models. Our proposed method consists of 3 parts such as situation dependent acoustic model, situation transition model and situation dependent language model. Due to these models, our proposed method *Situation Based Speech Recognition* enables incorporating baseball task-dependent knowledge such as a structure of baseball and an announcer's emotion into speech recognition method. For example, in conventional speech recognition, "Ball count two and two" could occur immediately after "Ball count one and one" due to speech recognition error. Following the baseball rule, it does not occur. In the proposed speech recognition, the probability from "ball: 1, strike: 1" to "ball: 2, strike: 2" i.e. $P(B2, S2|B1, S1)$ is set to be zero. Therefore, the proposed speech recognition can prevent an incorrect recognition. This method based on baseball situations is formalized in the framework of probability theory and implemented in the conventional speech decoding (Viterbi) algorithm.

2. Knowledge of Baseball Games

A content of a baseball game consists of a sequence of video data and speech data. We use a commentary speech on a radio. It has following features. First, it contains much more information about keywords than TV. Second, the radio speech is rather fast, noisy, emotional and disfluent. Therefor, it is a difficult task for speech recognition systems to recognize the radio speech with a high accuracy. Third, an announcer speaks in a situation of a baseball game. Hence commentary speech depends on a situation of a baseball game such as counting of inning, out, ball and strike. We define a sequence of these information as a *structure* of baseball games (shown in Figure 1). Last, an announcer is often excited when a situation is exciting. From above features, it can be thought that speech recognition should be performed using a situation of baseball game such as meta information and an announcer's emotion to improve speech recognition accuracy. To develop this kind of speech recognition, we propose the speech recognition method that estimates a word sequence as well as baseball situation sequence simultaneously. To find the correct sequence of the baseball game situations, the speech recognition accuracy, par-

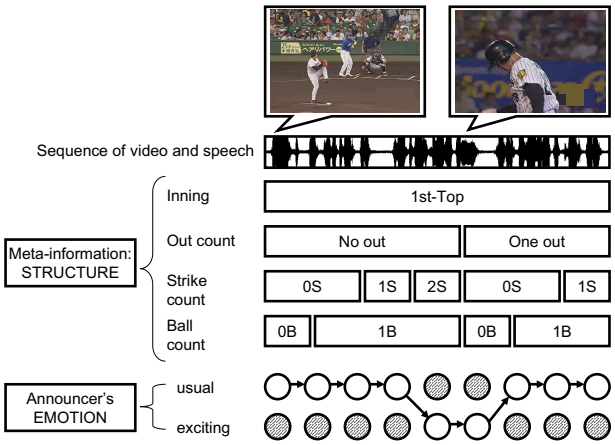

Figure 1: *The structure of a baseball game.*

ticularly the keywords accuracy is important, because the keywords are deeply related to the situation of a baseball game. In the past, to improve the performance of speech recognition in this task, we performed the acoustic model adaptation and the language model adaptation. Herewith, the performance was quite improved, but not enough to accomplish our goal. Now we notice that the utilization of heuristic rules related to the keywords plays an important role to perform structuring of a baseball game.

3. Situation Based Speech Recognition

In this section, we formalize the proposed speech recognition and compare it to the conventional speech recognition, through their formulation. Let $\mathbf{O} = \{O_1, \cdots, O_T\}$ and $\mathbf{W} = \{W_1, \cdots, W_N\}$ be a sequence of D-dimensional observed feature vectors and a word sequence respectively. The general problem of speech recognition is to find the most likely word sequence $\mathbf{W}$, given the sequence of observed feature vectors $\mathbf{O}$. But the goal of our proposed method "Situation Based Speech Recognition" is to find the most likely word sequence $\mathbf{W}$ and *situation sequence* $\mathbf{S} = \{S_1, \cdots, S_N\}$ *simultaneously*, given the sequence of observed feature vectors $\mathbf{O}$ as follows:

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{S}, \mathbf{W}|\mathbf{O}). \quad (1)$$

Eq.2 can be derived from Eq.1:

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{S}, \mathbf{W}) P(\mathbf{O}|\mathbf{W}, \mathbf{S}), \quad (2)$$

based on Bayesian theorem and $P(\mathbf{O})$ is omitted due to independence from $\mathbf{W}$. Here, the probability $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ is a situation dependent acoustic model. Moreover, we can derive the following equation.

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(S_1, \cdots, S_N, W_1, \cdots, W_N) \\ &= P(S_1)P(W_1|S_1) \\ &\times \prod_{i=2}^{N} P(S_i|S_1^{i-1}, W_1^{i-1}) P(W_i|S_1^i, W_1^{i-1}). \end{aligned} \quad (3)$$

Based on the following approximation,

- A situation depends only on a previous situation, a previous word and the word which makes transition probability highest in past M words.
- A word depends only on a present situation and a previous word.

we can simplify Eq.3 as follows:

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(S_1)P(W_1|S_1) \\ &\times \prod_{i=2}^{N} P(S_i|S_{i-1}, W_{i-1}, W_y) P(W_i|W_{i-1}, S_i), \end{aligned} \quad (4)$$

where

$$W_y = \underset{W_j=(W_{i-1}\cdots W_{i-M})}{\operatorname{argmax}} P(S_i|S_{i-1}, W_i, W_j). \quad (5)$$

Finally, the speech recognition with situation estimation is formalized as:

$$\begin{aligned} (\hat{\mathbf{S}}, \hat{\mathbf{W}}) &= \underset{\mathbf{S},\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}, \mathbf{S}) P(S_1) P(W_1|S_1) \\ &\times \prod_{i=2}^{N} P(S_i|S_{i-1}, W_{i-1}, W_y) P(W_i|W_{i-1}, S_i). \end{aligned} \quad (6)$$

Note that $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ is an acoustic model depending on a situation, $P(S_i|S_{i-1}, W_{i-1}, W_y)$ is a situation transition probability and $P(W_i|W_{i-1}, S_i)$ is a bi-gram probability depending on a situation.

Now we compare the proposed method shown in Eq.6 to the conventional method formalized as:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W}) \prod_{i=1}^{N} P(W_i|W_{i-1}). \quad (7)$$

Firstly, the acoustic probabilities $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ is different from the conventional method. Due to this acoustic model, it can be expected that multiple acoustical features such as normal, exciting, sad or other emotions can be represented. Especially in this paper, we categorize the speeches into two groups: "normal" and "exciting" speeches. Secondly, there is the situation transition probability in the proposed method, not in the conventional method. This model plays a role to predict next situation from speech recognition results and to prevent a wrong situation transition not to occur under the baseball rules. Finally, the bi-gram probability depending on baseball game situation $P(W_i|W_{i-1}, S_i)$ in the proposed method is similar to the bi-gram probability $P(W_i|W_{i-1})$ in the conventional method. However in the proposed method, it can be estimated depending on a situation. Namely the word probability can be regarded as depending on a situation.

Next, we describe how to learn the stochastic models in the proposed speech recognition such as the acoustic model depending on the situation $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$, the situation transition probability $P(S_i|S_{i-1}, W_{i-1}, W_y)$ and the bi-gram probability depending on the baseball game situation $P(W_i|W_{i-1}, S_i)$.

4. Learning Stochastic Models

4.1. Situation Dependent Acoustic Model

We use two HMMs to represent a situation dependent acoustic model $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$. The recognition is performed by using both HMMs simultaneously and select a word with the higher likelihood. The one HMM represents a normal emotion and the

other HMM represents excited emotion. Each HMM is created by the adaptation techniques of acoustic model using speech corpus for each emotion. The adaptation is performed by the speech which specked by same speaker as the one for the test set. We performed the adaptation as follows,

- First, excited time sections are detected by human listening as adaptation corpus.
- Next, speech data and the corresponding text data at the excited time sections are separated from normal emotion time section.
- Next, model parameters are transformed by MLLR (Maximum Likelihood Linear Regression) using supervising text.
- Last, MAP (Maximum A Posteriori) estimation is performed using the transformed parameters.

Here, the adaptation at exciting time sections is performed after the adaptation at the normal emotion time sections because amounts of exciting time sections are small.

4.2. Situation Transition Model

This model has two meanings as shown in figure 2. The one is the representation of a simple baseball rule network. In other words, this model has an ability that can prevent a wrong situation transition such as from "ball count two and one" to "ball count three and two". To actualize this ability, the situation transition probability $P(S_i|S_{i-1}, W_{i-1}, W_y)$ which is not allowed by the baseball rules is set to be zero independently of words.

The other is the representation of a situation prediction from speech recognition results. For example, suppose that present situation is ball count two balls and one strike. When the recognition result of an announcer's utterance is "Pitch and strike" or "Hit, foul ball", it is expected that a situation will transit to ball count two balls and two strikes. This can be considered a kind of topic prediction problems i.e. a situation is transited depending on a topic of an announcer's utterance. In this paper, a topic is predicted by only words approximately. If we use N words, the data sparseness problem is occurred. But if we use only a word, the wrong prediction is occurred. For example, it takes *strike* of the utterance "ball count two balls and one *strike*" for "pitch and *strike*". Hence we have decided to use two words.

A situation transition probability $P(S_i|S_{i-1}, W_{i-1}, W_y)$ is learned by the following steps.

- Give the situation labels into transcription text corpus by human hand.
- Perform the morpheme analysis.
- Compute probability by following the equation.

$$P(s_i|s_{i-1}, w_i, w_y) = \frac{N(w_i, w_y, s_i|s_{i-1})}{N(w_i, w_y)}, \quad (8)$$

where $N(x, y)$ is frequency of co-occurrence x and y.

4.3. Situation Dependent Language Model

The bi-gram probability depending on the situation $P(W_i|W_{i-1}, S_i)$ can be obtained through learning. However it invokes the data sparseness problems so that we focus on the following three types of commentaries on a baseball game as shown in Table 1.

Type (i) is related to a situation. If an utterance of this type occurs, it leads to a situation transition. But note that utterance of this type is appeared only in an appropriate situation. For example, utterance of "Swinging strikeout." is uttered only in strike count two.

Type (ii) is for explaining a baseball game situation. For example, if utterance of "Count is two and two." is spoken, it indicates high probability that a situation is in ball count two balls and two strikes now. Our experience shows that utterances of this type can correct a wrong situation prediction.

Type (iii) is not related to a baseball situation. An utterance in this type doesn't lead to situation transition nor explain a situation. So, it can be thought that a sequence of words $\mathbf{W}$ doesn't depend on a sequence of situation $\mathbf{S}$. Therefore, in this type, $P(W_i|W_{i-1}, S_i)$ can be reduced to $P(W_i|W_{i-1})$. This equation is same as the bi-gram probability in the conventional speech recognition.

For type (i) and (ii) commentaries, we manually marked the baseball game situation to the training data, and computed the bi-gram probability statistically from this training data. In the type (iii), we used a bi-gram same as the conventional speech recognition.

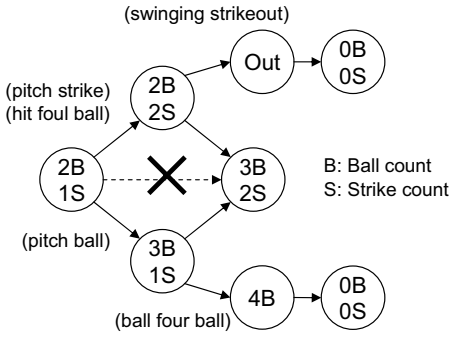

Figure 2: *An image of situation transition model.*

Table 1: *Types of commentaries and the examples.*

	Types	Examples
(i)	Related to a situation transition	"Swinging strikeout." "Ball, four balls."
(ii)	Explain a situation	"Count is two and two." "Two out and two on."
(iii)	No relation with a situation	"A bird flies..." "It is cloudy with a chance of rain."

5. Experiments

5.1. Experimental Conditions

The proposed speech recognition method described in Section 3 enables incorporating baseball dependent knowledge into speech recognition and also enables visualizing a sequence of baseball game situations and time sections of speaker's emotion. We carried out experiments to prove the effectiveness of the proposed speech recognition.

The experimental conditions are summarized in Table 2. In order to improve the speech recognition accuracy for a commentary on a baseball, we employed the acoustic and language model adaptation. In the acoustic model, the training data

for a baseline consisted of about 200,000 Japanese sentences (200 hours) spoken by 200 males in CSJ (Corpus of Spontaneous Japanese)[3]. The adaptation was performed based on the method derived in Section 4.1. In the language model, the training data for a baseline consisted of 570,000 morphemes collected from web pages about baseball. This was named "Web text corpus". We also made "Dictated text corpus" that was manually transcribed from commentary speech on a baseball game. Then, we merged these two corpora into one corpus named "Merged text corpus". Finally, we further merged "Dictated text corpus" and "Merged text corpus" with weighting under condition of minimum perplexity. We selected the keywords deeply related to the structure of a baseball game as shown in Table 3. The number of baseball game situations was 72 (3 strikes, 4 balls, 3 outs and 2 emotions). Experiments were carried out under these conditions, using decoder based on ML back-off[4].

Table 2: *Experimental conditions*

Sampling rate/Quantization	16 kHz / 16 bit
Feature vector	26 - order MFCC
Window	Hamming
Frame size/shift	20/10ms
# of phoneme categories	244 syllable
# of mixtures	32
# of states (Vowel)	5 states and 3 loops (Left to Right)
# of states (Consonant+Vowel)	7 states and 5 loops (Left to Right)

Table 3: *Keywords*

Strike, Ball, Four balls (Base on Balls in English)
Missed swing, Strike out, Foul, Out

5.2. Experimental Results

Table 4 shows the experimental results. "Baseline" indicates a result by conventional speech recognition using the acoustic model and language model adapted. "Proposed method" is a result by our method incorporating baseball task dependent knowledge into speech recognition. Note that "correct rate of structuring" is a percentage of the correctly recognized structure (inning, out count, strike count, ball count) to the total number of structure and "correct rate of detecting excited scenes" is a percentage of the correctly recognized time sections in excited speech. We can improve the keyword accuracy by 2.3%. This is mainly because the proposed method prevented incorrect words based on baseball dependent knowledge. Figure 3 shows the prospect of this method. For example, there is an utterance such as "... foul ball, and strikeout in next pitch". In conventional speech recognition, it misrecognized "four balls (means base on balls in English)" instead of "foul ball" due to very similar pronunciation. But the strikeout should not occur immediately after four balls because four balls sets the strike count to zero. The strikeout can occur under the condition of strike count two. The proposed speech recognition can estimate a sequence of game situations and correctly recognize "foul ball, and strikeout". We confirmed 73.3 % of the correct rate of structuring and 75.0% of the correct rate of detecting excited scenes.

Table 4: *Experimental results*

	Baseline	Proposed method
Keyword accuracy	66.8 %	69.1%
Correct rate of structuring	-	73.3%
Correct rate of detecting excited scenes	-	75.0%

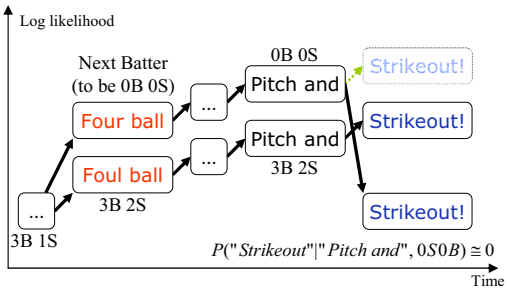

Figure 3: *The prospect of how to prevent an incorrect recognition.*

6. Summary

In this paper, we described how to structure the commentary speech by incorporating baseball task dependent knowledge and proposed the situation based speech recognition method. It can be thought that the proposed speech recognition is a sort of the information integration. The experimental result showed that the proposed method improved the 2.3% of keyword accuracy, and herewith achieved the 73.3% of the correct rate of structuring and 75.0% of the correct rate of detecting excited scenes.

7. References

[1] A. Sako, Y. Ariki: "Structuring Baseball Live Games Based on Speech Recognition Using Task Dependent Knowledge and Emotion State Recognition", in ICASSP 2005, pp. 1049-1052, 2005.

[2] Y. Ariki, T. Shigemori, T. Kaneko, J. Ogata and M. Fujimoto: "Live Speech Recognition in Sports Games by Adaptation of Acoustic Model and Language Model", in Eurospeech2003, pp.1453-1456, 2003-09.

[3] S. Furui, K. Maekawa, H. Isahara: "Spontaneous Speech: Corpus and Processing Technology", The Corpus of Spontaneous Japanese, pp.1-6, 2002-2.

[4] J. Ogata, Y. Ariki: "An Efficient Lexical Tree Search for Large Vocabulary Continuous Speech Recognition", Proc. of the Sixth Int'l Conf. on Spoken Language Processing(ICSLP'00), Vol.II, pp.967-970 (Oct. 2000).

[5] C. Popovic and P. Baggia: "Specialized language models using dialogue predictions", in ICASSP 1997, pp 423-426, 1997.

[6] F. Wessel and A. Baader: "Robust dialogue-state dependent language modeling using leaving-one-out", in ICASSP 1999, pp741-744, 1999.

[7] Wei Xu and Alex Rudnicki: "Language modeling for dialogue systems", in ICSLP 2000, pp 118-121, 2000.